



# Training Data Lab

Aplicaciones de data mining, modelamiento y machine learning  
para ciencias sociales

---

**Bastián González-Bustamante**

University of Oxford

Universidad de Santiago de Chile

✉ [bastian.gonzalezbustamante@politics.ox.ac.uk](mailto:bastian.gonzalezbustamante@politics.ox.ac.uk)

Presentación preparada para el primer taller del laboratorio de datos  
Training Data Lab, 20 de abril de 2021 (**versión en progreso**)

# Hoja de ruta

1. Introducción
2. Proyectos realizados
3. Proyectos en curso
4. Próximos proyectos
5. Tufte Working Papers
6. Lineamientos de colaboración



# Introducción

---

# Introducción

© 2020 **Training Data Lab** es un grupo de investigación se enfoca en aplicaciones de ciencia de datos aplicada a ciencias sociales en tres áreas interconectadas: **minería de datos, modelamiento econométrico y aprendizaje automático**. Por una parte, buscamos recoger datos con técnicas de minería para elaborar modelos econométricos con técnicas observacionales o de emparejamiento.

Por otro lado, nos enfocamos en entrenar modelos con técnicas de aprendizaje automático y profundo etiquetando conjuntos de datos para diferentes proyectos. Lo anterior, nos permite clasificar datos no codificados usando nuestros modelos entrenados incorporando validación humana en el flujo de trabajo, lo que mejora la inteligencia artificial en los procesos de aprendizaje.



Universiteit  
Leiden

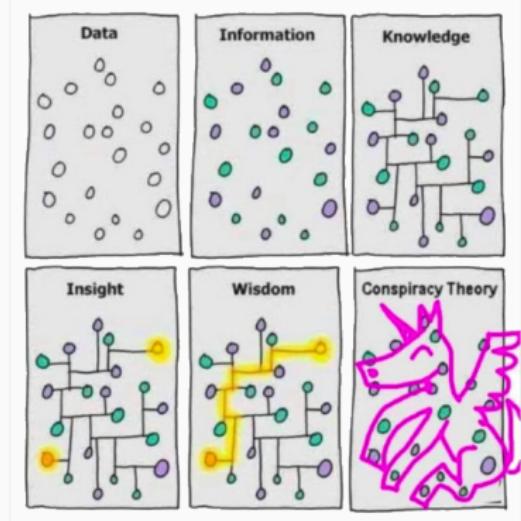
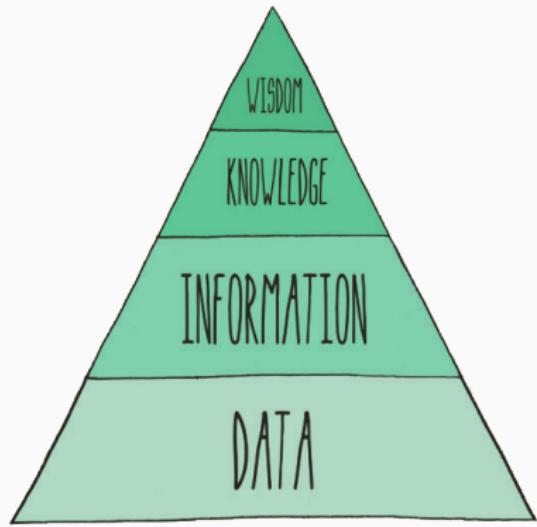


UNIVERSIDAD  
CATÓLICA DE  
TEMUCO



UNIVERSIDAD  
MAYOR

# Introducción



## Proyectos realizados

---

# Rastreador Online COVID-19

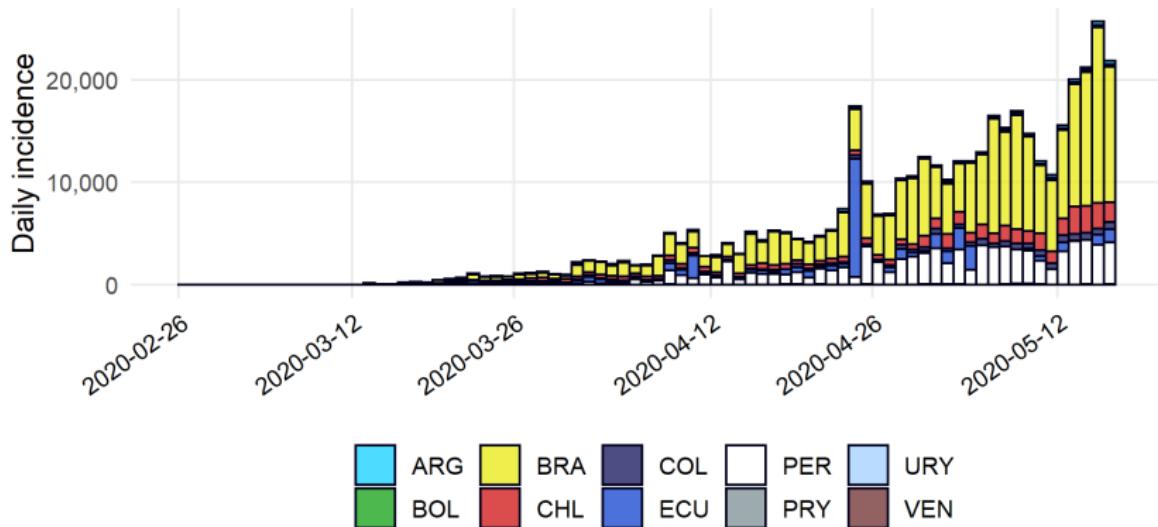
-  Bastián González-Bustamante (responsable)
-  [bgonzalezbustamante.github.io/COVID-19-South-America](http://bgonzalezbustamante.github.io/COVID-19-South-America)

**COVID-19 in South America Tracker.** Rastreador online para Sudamérica desplegado entre el 11 de marzo y mediados de mayo de 2020 con datos de JHU. Este tracker presentaba periódicamente las curvas epidémicas en función de la incidencia en los distintos países de la región, el R estimado, la distribución de intervalo de serie explorada (SI) y simulaciones de incidencia futura. Además, se ofrecían comparaciones con algunos países europeos.

**Early Government Responses to COVID-19 in South America.** Artículo publicado en World Development (González-Bustamante, 2021), código en R disponible en el repositorio y archivo completo de replicación en Elsevier que integra datos de JHU, Oxford, PAHO, V-Dem y WB.

# Rastreador Online COVID-19

## COVID-19 - Coronavirus Epidemic Curve in South America

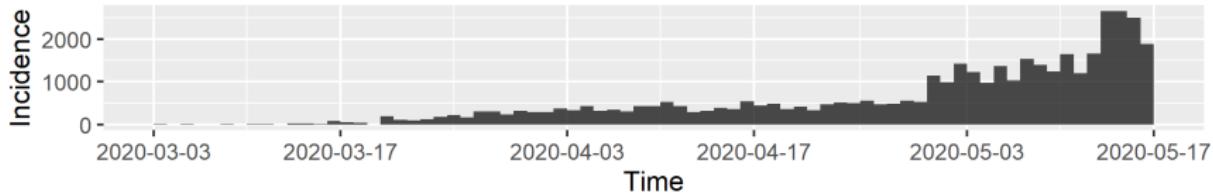


Data up to 16 May - DOI: 10.17605/OSF.IO/Y6C7Z

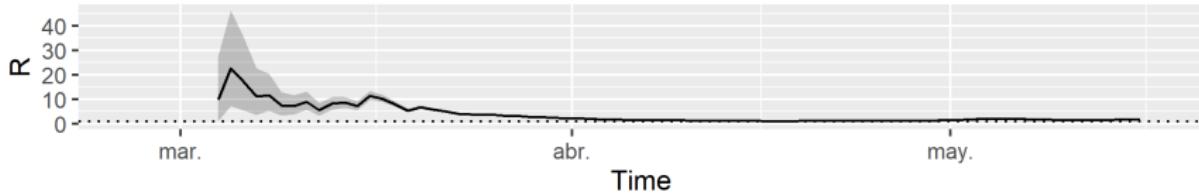
Note: There is inconsistency on 12 April data in the Uruguayan case.  
As well as on 7, 8, and 9 May data in the Ecuadorian case.

# Rastreador Online COVID-19

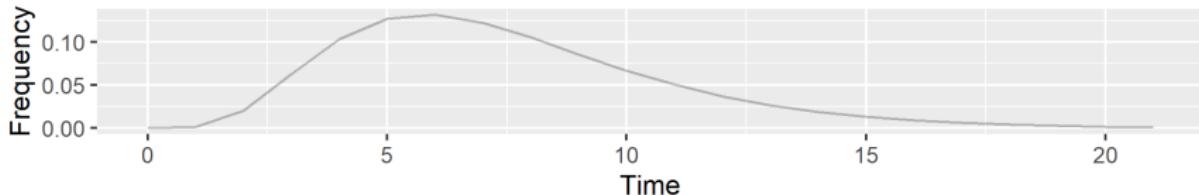
Epidemic curve



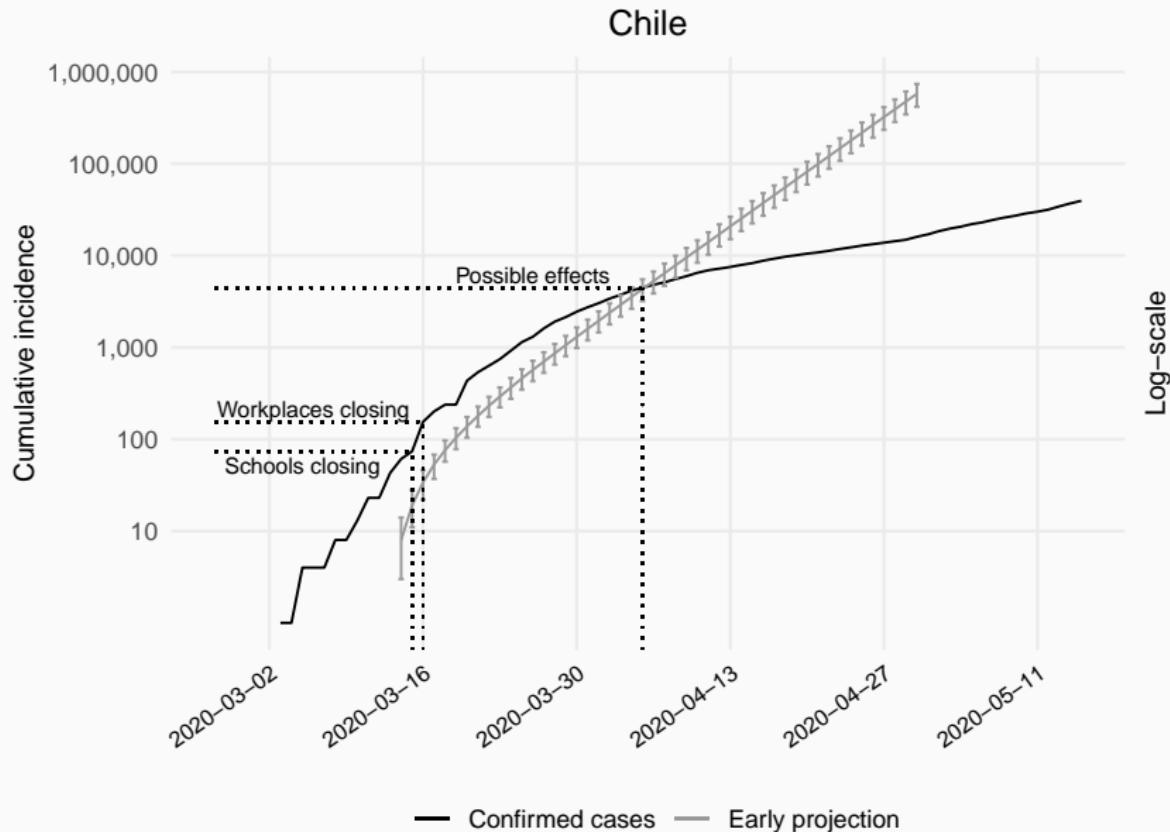
Estimated R



Explored SI distribution



# Rastreador Online COVID-19



# Algoritmo OCR para servicio civil chileno

 Bastián González-Bustamante, Matías Astete y Berenice Orvenes  
(responsables)

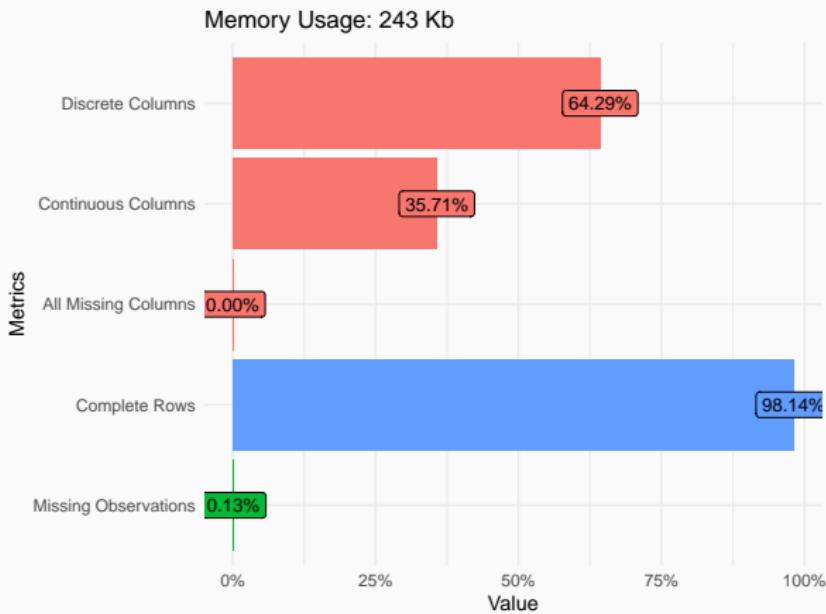
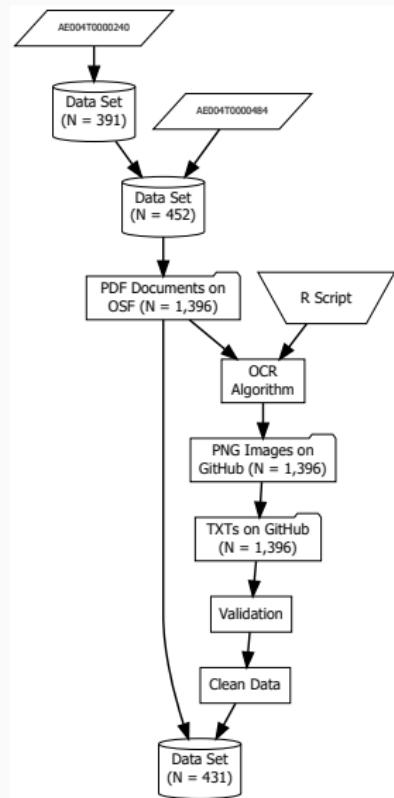
 DOI: [10.17605/OSF.IO/WBF6M](https://doi.org/10.17605/OSF.IO/WBF6M)

 Pronto disponible en [training-datalab.com](https://training-datalab.com)

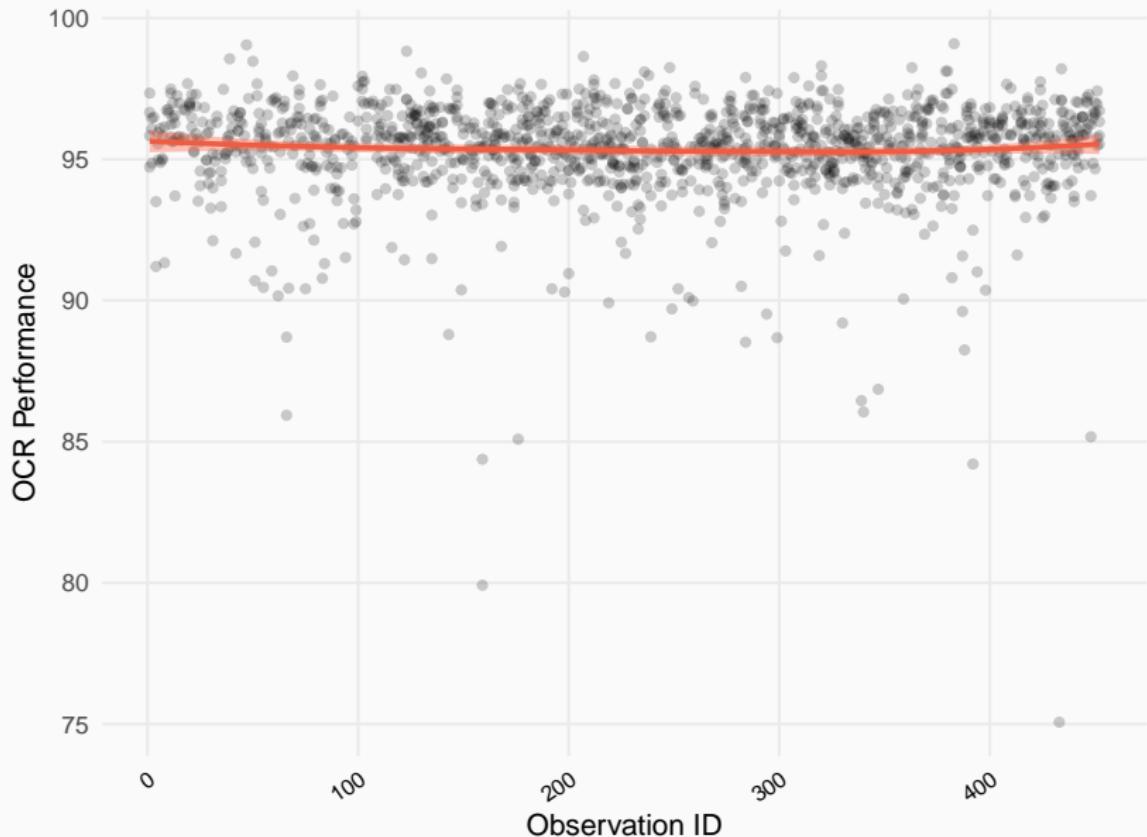
**A Novel Dataset on Members of the Chilean Civil Service.** Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

 Revisar el documento de trabajo en este repositorio privado (no distribuir por favor).

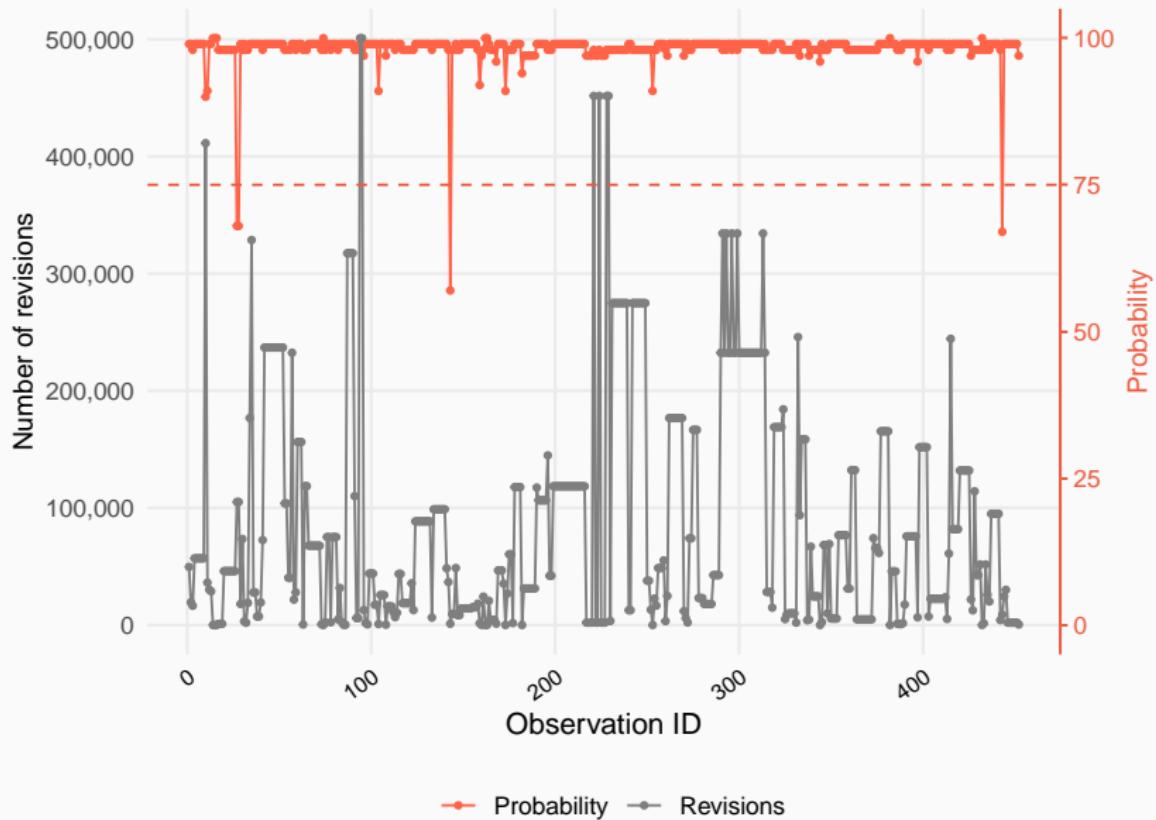
# Algoritmo OCR para servicio civil chileno



# Algoritmo OCR para servicio civil chileno



# Algoritmo OCR para servicio civil chileno



## Proyectos en curso

---

# Minería de datos en Google Scholar

-  Bastián González-Bustamante (responsable)
-  Alejandro Olivares y Carla Cisternas (colaboradores)
-  [bgonzalezbustamante.com/cps-ranking](http://bgonzalezbustamante.com/cps-ranking)

**Chilean Political Science Ranking (CPS-Ranking).** Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

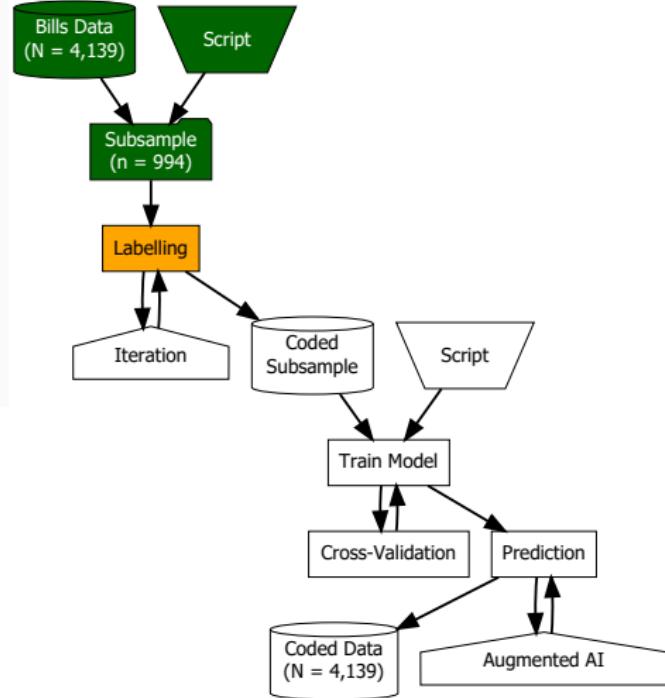
\*\* INSERT FIGURE \*\*

# Algoritmo clasificador para mociones legislativas

-  Carla Cisternas y Bastián González-Bustamante (responsables)
-  Diego Aguilar (colaborador)
-  Estamos reclutando colaboradores y ayudantes
-  [training-datalab.com/projects/chilean-congress-bills](http://training-datalab.com/projects/chilean-congress-bills)

**Training Data on Chilean Congress Bills.** Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

# Algoritmo clasificador para mociones del Congreso chileno



# Deep learning para clasificar publicaciones

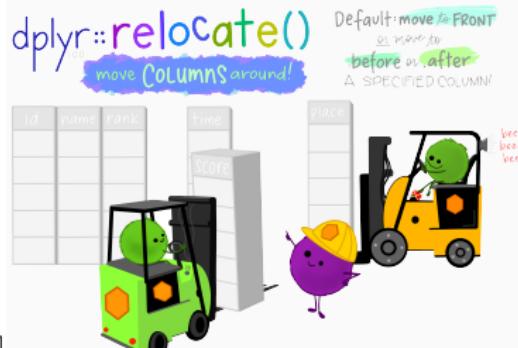
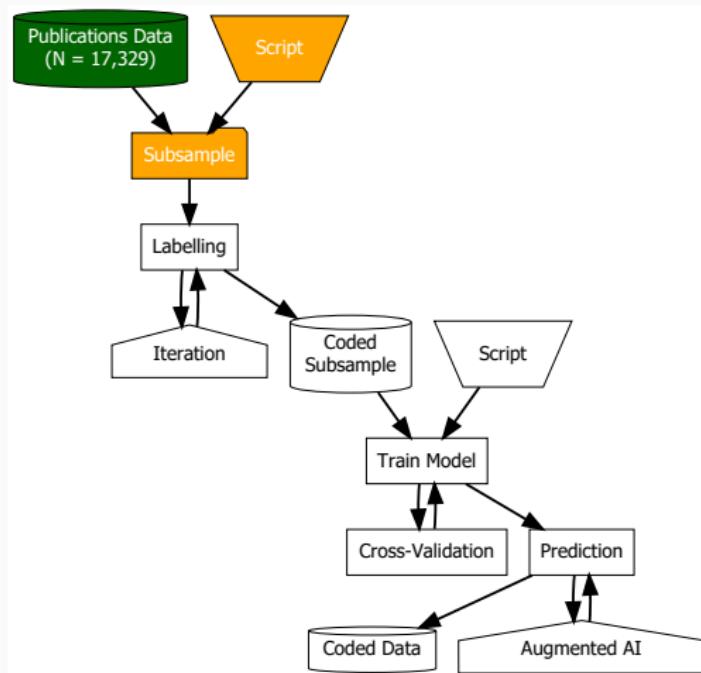
 Bastián González-Bustamante, Alejandro Olivares y Carla Cisternas  
(responsables)

 [training-datalab.com/projects/political-science-publications](http://training-datalab.com/projects/political-science-publications)

Algoritmos de aprendizaje profundo para clasificar la producción científica: Evidencia de la ciencia política en español. Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

 Revisar el abstract en este repositorio ([no distribuir por favor](#)).

# Deep learning para clasificar publicaciones



## Próximos proyectos

---

# Proyecto VIP de machine learning

-  Carla Cisternas y Francisco Castañeda (responsables)
-  Bastián González-Bustamante, Rodrigo Cuevas, Alejandro Olivares y Mariana Ardiles (colaboradores)
-  Pronto disponible en [training-datalab.com](http://training-datalab.com)

## Aplicaciones de Machine Learning en políticas públicas y economía.

Este proyecto VIP busca conformar un equipo multidisciplinario de investigación verticalmente integrado para realizar diferentes aplicaciones de aprendizaje automático en temas de políticas públicas y economía. En esta primera versión, nos centramos en las dinámicas de desinformación en temas económicos y de interés público.

Posteriormente, en futuras versiones, esperamos abordar y analizar otros fenómenos relacionados con la formulación de políticas públicas sectoriales y regulación de mercados.

# Proyecto VIP de machine learning



## APLICACIONES DE MACHINE LEARNING EN POLÍTICAS PÚBLICAS Y ECONOMÍA

# Scraper twConstitution

-  Bastián González-Bustamante (responsable)
-  Estamos reclutando colaboradores y ayudantes
-  [bgonzalezbustamante.github.io/twConstitution](https://bgonzalezbustamante.github.io/twConstitution)
-  Pronto disponible en [training-datalab.com](https://training-datalab.com)

**Twitter Online Tracker of the Chilean Referendum for a New Constitution.** Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

\*\* INSERT FIGURE \*\*

# Tufte Working Papers

---

# Tufte Working Papers

-  Bastián González-Bustamante (editor)
-  Elinor Luco (asistente editorial)
-  Pronto disponible en [training-datalab.com](http://training-datalab.com)

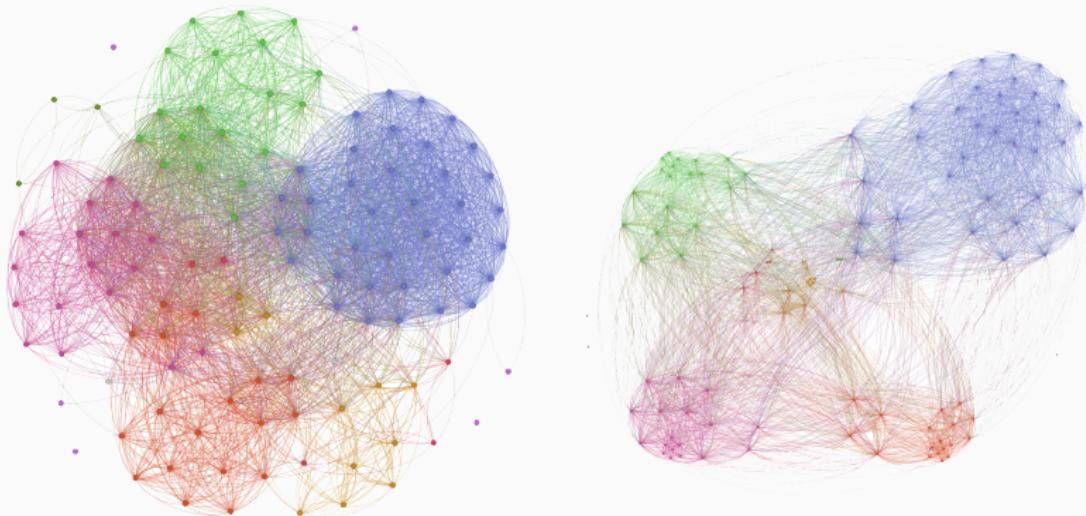
© 2020 **Tufte Working Papers (ISSN 2735-6043)** es una publicación continua basada en Training Data Lab con el apoyo logístico de la Universidad de Santiago de Chile. Esta serie de documentos promueve el debate en ciencias sociales, especialmente en temas relacionados con la ciencia política y las políticas públicas. La serie incluye trabajos inéditos y versiones revisadas de publicaciones previas que proponen **técnicas de investigación innovadoras u ofrecen información empírica novedosa**. Los trabajos se publican en español e inglés, aunque se privilegian las publicaciones en español con el fin de apoyar la difusión del conocimiento de acceso abierto en Iberoamérica.

Esta serie es editada en **Tufte-LaTeX**, una plantilla LaTeX inspirada por Edward R. Tufte. LaTeX es un software libre que permite la composición de textos con alta calidad tipográfica. Por otro lado, cuenta con una **política ética y normas de estilo** basadas en diversas declaraciones e iniciativas (Singapur, COPE, Budapest, Bethesda y Berlín). En consecuencia, los trabajos se licencian bajo Creative Commons.

La serie utiliza la **taxonomía CRediT** para identificar las contribuciones exactas de cada investigador en caso de coautorías o asistencias de investigación.

Se utiliza una **revisión abierta** con uno o dos expertos en la que se divultan las identidades de los autores y árbitros.

Aplicación de ForceAtlas2, un algoritmo de diseño gráfico continuo, para el estudio de las élites. Working paper enfocado en análisis de redes sociales y su visualización ([González-Bustamante y Cisternas, 2020](#)).



## **Lineamientos de colaboración**

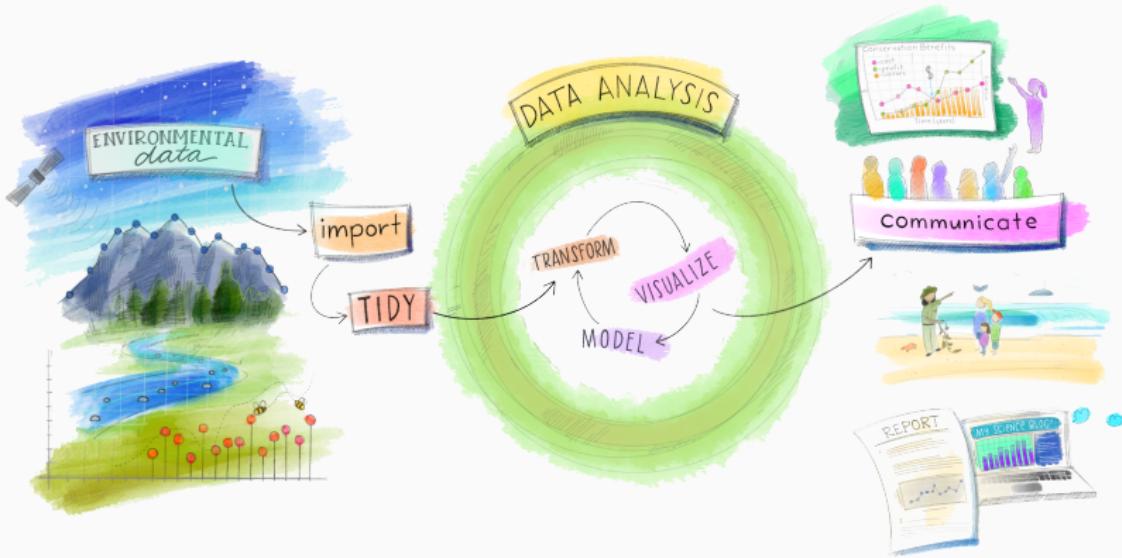
---

# Lineamientos de colaboración

**Taxonomía CRediT.** Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.



# Lineamientos de colaboración



**Solo un poco de información adicional...**

# Contacto

## Bastián González-Bustamante

DPhil (PhD) Researcher

Department of Politics and International Relations  
& St Hilda's College  
University of Oxford

📍 St Hilda's College, Cowley Place, Oxford OX4 1DY

✉️ [bastian.gonzalezbustamante@politics.ox.ac.uk](mailto:bastian.gonzalezbustamante@politics.ox.ac.uk)

🏠 <https://bgonzalezbustamante.com>

## Profesor Instructor

Departamento de Gestión y Políticas Públicas

Facultad de Administración y Economía

Universidad de Santiago de Chile

📍 Av. Lib. B. O'Higgins 3363, Estación Central, Santiago

✉️ [bastian.gonzalez.b@usach.cl](mailto:bastian.gonzalez.b@usach.cl)

Presentación compilada con **LATEX** y algunos 

 Descargar la versión más reciente desde [GitHub](#)

 Descargar el video de la presentación desde [Dropbox](#)

 Artwork utilizado disponible en [GitHub](#)

**Muchas gracias por su atención**

**[training-datalab.com](http://training-datalab.com)**